REMARKS/ARGUMENTS

Applicant thanks the Examiner for the allowance of claims 42-56 and finding claims 4-5, 7, 10, 16-24, 29-31, 35, and 37-41 to be allowable if rewritten as independent claims.

The amendment of claims 11, 13, and 25 does not warrant further searching of the prior art. Independent claim 11 has been amended to include the limitations of dependent claim 15. Independent claim 25 has been amended to include the limitations of dependent claim 28. Finally, dependent claim 13 has been amended to correct a typographical error.

The Examiner has rejected claims 1-3, 6, 8-9, 13, 27, 34, and 36 under 35 U.S.C.§103(a) as being unpatentable over Colby et al. (U.S. 6,862,624) in view of Gupta et al. (U.S. 6,718,387) and claims 11-12, 14-15, 25-26, 28, and 32-33 under 35 U.S.C.§102(e) as being anticipated by Colby et al.

The Rejection of Claims 1-3, 6, 8-9, 11-15, 25-28, and 32-34

The cited references fail to teach or suggest at least the italicized features of independent claims 1, 11, 25, and 32:

> 1. A network switch for switching transaction requests among a plurality of servers, the network switch being positioned between the plurality of servers and at least one client, comprising:
>
> a parser operable to parse transaction requests to locate one or more selected fields;
>
> a router operable to forward at least portions of the transaction requests to respective servers in the plurality of servers and transaction responses of the respective servers to the transaction requests to respective clients; and
>
> *a tag generator operable to generate a tag associated with a selected server in the plurality of servers and include the tag in a transaction response received from the selected server, the transaction response comprising information requested by a transaction request and a cookie generated by the selected server, whereby, when a subsequent transaction request is received from the client corresponding to the tagged transaction request, the subsequent*

transaction request includes the tag and the cookie and, based on the tag, the router forwards the subsequent transaction request to the selected server.

11. A method for switching transaction requests, comprising:

receiving, from a first source, a transaction response associated with first source, the transaction response corresponding to at least a first transaction request;

parsing the transaction response to locate at least a first field;

determining a first tag identifying the first source;

appending the first tag to the first field in the transaction response;

reassembling the transaction response;

forwarding the transaction response to a destination identified by the transaction response, wherein the first source is a first server in a plurality of servers and the destination is a client;

receiving the transaction response after the forwarding step;

storing the first tag in the client's memory;

forwarding a second transaction request to an address associated with the first server, *the second transaction request including the first tag*;

receiving the second transaction request from the client;

parsing for the first field in the second transaction request; and

forwarding the second transaction request to the first server based on the first tag.

25. A system for switching transaction requests among a plurality of servers, comprising:

an input port for receiving, from a first server in the plurality of servers, a transaction response of the first server, the transaction response corresponding to at least a first transaction request;

means for parsing the transaction response to locate at least a first field;

means for determining a first tag identifying the first server;

means for appending the first tag to the first field in the transaction response;

means for reassembling the transaction response;

means for forwarding the transaction response to a client identified by the transaction response;

a second input port for receiving the transaction response from the forwarding means;

means for storing the first tag in the client's memory; and

means for forwarding a second transaction request to an address associated with the first server, the second transaction request including the first tag, wherein each server in the plurality of servers has a unique identifier and the first tag is based on the unique identifier associated with the first server.

32. A system, comprising:
a communications network;
a plurality of replicated servers connected to the network, *all of the replicated servers having a same network address and all of the replicated servers serving the same replicated information,* each of the replicated servers being configured to receive a first transaction request associated with an individual transaction and to provide a response to the first transaction request, *the response including a first tag that corresponds to the transaction, the first tag being a cookie generated by a first replicated server*; and
a network switch connecting the replicated servers to the network, the network switch being configured to generate a second tag associated with the first replicated server, to append the second tag to the first tag in the response, and to direct to the first replicated server subsequently received transaction requests including the first and second tags.

In one embodiment, the present invention is directed to a pipelining technique for use in a server farm. A server-specific tag, other than and different from a cookie, is appended to outgoing response packets to a client. When further request packets are received from the client, the server-specific tag attached to the request packets permits a flow switch to forward the response quickly to the appropriate server with a minimum degree of processing overhead. Parsing and processing of the cookie require much more processing resources than processing of the much smaller server-specific tag.

Colby et al.

Colby et al. is directed to a content-aware flow switch that intercepts a client content request in an IP network and transparently directs the content request to a best-fit server. The flow switch 110 "front-ends" (i.e., intercepts all packets received from and transmitted by) a set of local web servers 1000a-c constituting a web server farm 150. The best-fit server is chosen by the flow switch 110 based on the type of content requested, the quality of service or QoS requirements implied by the content request, the degree of load on available servers, network congestion information, and the proximity of the client to available servers. The flow switch 110 detects client-server flows based on the arrival of TCP SYNs and/or HTTP GETs from the client. The flow switch 110 implicitly deduces the QoS requirements of a flow based on the content of the flow. The flow switch 110 also provides the functionality of multiple physical web servers on a single web server in a way that is transparent to the client, through the use of virtual web hosts and flow pipes.

The flow pipes are logical pipes through which all flows between virtual web hosts and clients travel. The pipes are implemented using QoS tags. A QoS tag is created from the egress port of the flow switch to which the candidate server is connected, the ingress port of the flow switch at which the content request arrived, and other information from the candidate server record. The tag encapsulates information about the deduced QoS requirements of an existing or requested flow. (Col. 14, lines 52-61.) Such information includes the minimum requested bandwidth requirement of the requested content, and buffer requirements. (Col. 15, lines 53-55.) Two separate TCP sessions exist, one between the client and flow switch, and the other between the flow switch and the best-fit server. "As such, the IP, TCP, and possible content headers on packets moving bidirectionally between the client and server are modified as necessary as they traverse the content-aware flow switch 110." (Col. 16, lines 20-26.)

Colby et al. fails to teach what is tagged by the QoS tag. Is it a data structure or a packet? Is it a packet exchanged between the client and flow switch and/or between the flow switch and best-fit server?

Moreover, Colby et al. fails to teach the use of a tag that identifies uniquely a server assigned to service the content requests of a specific client. For this teaching, the Examiner relies on Gupta et al.

Gupta et al.

Gupta et al. is directed to a method for load balancing including creating a network, having a plurality of servers, to service a single multicast address using a source specific to join, where the source specific join allows each of the servers to specify a source internet protocol address range that each of the servers services. The method further includes reallocating the source internet protocol address range specified for one or more of the servers using a load balancing policy and control multicast channel while one or more of the servers is handling communications.

Gupta et al. discloses a process of "tag switching". The concept is discussed with reference to Figs. 17 and 18. Each server sets up a tagged switched tree for routing in the network. When a virtual IP router gets a tagless packet, it selects a server and forwards the tagless packet to the selected server. "The virtual IP router then informs the upstream router 1710 (and this router can in turn inform some or all of its upstream routers) to mark all packets from the user with the tag ID of the designated server 1820." (Col. 9, lines 56-60.) This is typically done by forwarding the packet to the actual IP address of a selected server and informing the upstream routers of the actual IP address to which subsequent packets should be directed. The upstream routers than mark packets for that connection with an appropriate tag ID. (Col. 2, line 62 - col. 3, line 2.) At col. 10, lines 5-24, an embodiment is discussed in which the client sends the tag ID and a cookie to the designated server.

Gupta et al. does not use the tag ID to uniquely identify an information server but rather a communication path. According to Gupta et al., the Tag Information Base or TIB is populated "with incoming and outgoing tags for all the routes it can access, so that all packets can be forwarded by simple label swapping." (Col. 9, lines 28-32.) It is further stated that "each server will create a tag switched path for each connection that would be lost to the server during reallocation (1600) . . . However, old connections *with a tagged path* continue until a convenient closing point and then the tagged path will be broken down (1630)." (*Id.* at lines 38-45 (emphasis supplied).)

Moreover, the Examiner's combination of the references is flawed as the references teach away from one another. Colby et al. uses tags to control QoS for a session. It relies on a different mechanism for selecting and routing packets to a server. Colby et al. switches packets based on a combination of source and destination IP addresses, transport layer protocol, and transport layer source and destination port numbers. (Col. 6, lines 16-19.) Gupta et al. does not teach central control of server selection in a flow switch. Each of a distributed number of routers selects a server based on the destination address. Unlike the flow switch of Colby et al., each of the information servers in Gupta et al. specifies a source internet protocol address range that it will serve. When a server gets overloaded, it forwards all new connection requests to a less loaded server in accordance with distribution policy.

<u>The Rejection of Claim 36</u>

The Examiner rejects independent claim 36 under 35 U.S.C.§103(a) as being unpatentable over Colby et al. in view of Gupta et al.

Applicant respectfully traverses the Examiner's rejection. The cited references fail to teach or suggest at least the following italicized features of independent claim 36:

> 36. A method for providing information from a server to a client, comprising:
> receiving a first transaction request requesting first information, the first information referencing at least second and third information;
> retrieving the first information;
> providing the first information to the client;
> *determining which of the second and third information has been more frequently requested by clients during a first selected time interval;*
> *retrieving the more frequently requested of the second and third information and/or an address associated therewith;*
> *thereafter receiving a second transaction request from the client requesting the more frequently requested of the second and third information;* and
> providing the more requested of the second and third information to the client.

As noted previously, neither reference is directed to predicting the set or collection of information likely to be requested by a client let alone predicting the information based on the historic frequencies with which sets of information have been requested by other clients.

Accordingly, the pending claims are allowable.

The dependent claims provide further bases for allowance.

By way of example, the Examiner has found numerous dependent claims to be allowable if restated as independent claims.

Based upon the foregoing, Applicants believe that all pending claims are in condition for allowance and such disposition is respectfully requested. In the event that a telephone conversation would further prosecution and/or expedite allowance, the Examiner is invited to contact the undersigned.

Respectfully submitted,

SHERIDAN ROSS P.C.

By: Douglas Swartz
Douglas W. Swartz
Registration No. 37,739
1560 Broadway, Suite 1200
Denver, Colorado 80202-5141
(303) 863-9700

Date: July 12, 2005